DeepVoxels: Learning Persistent 3D Feature Embeddings
— Supplemental Document —

Vincent Sitzmann[1], Justus Thies[2], Felix Heide[3],
Matthias Nießner[2], Gordon Wetzstein[1], Michael Zollhöfer[1]

[1]Stanford University, [2]Technical University of Munich, [3]Princeton University

vsitzmann.github.io/deepvoxels/

1. Results on two Additional Synthetic Scenes

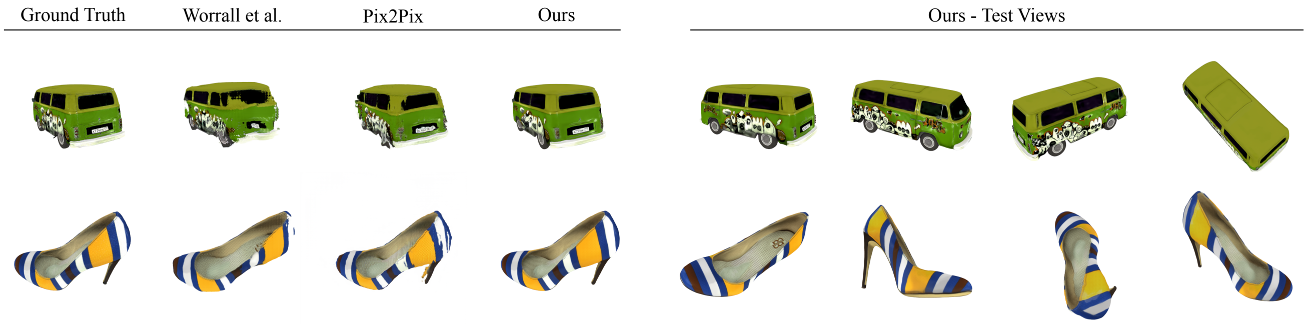

Figure 1: Qualitative results on two additional scenes.

	Bus	Shoe
	PSNR / SSIM	PSNR / SSIM
Nearest Neighbor	17.96 / 0.89	17.49 / 0.88
Tatarchenko et al. [2]	22.58 / 0.94	20.00 / 0.91
Worrall et al. [3]	19.30 / 0.91	20.34 / 0.91
Pix2Pix (Isola et al.) [1]	24.41 / 0.95	23.45 / 0.93
Ours	**31.78 / 0.98**	**33.70 / 0.98**

Table 1: Quantitative comparison to four baselines on two additional scenes. Our approach obtains the best results in terms of PSNR and SSIM on all objects. See Fig. 1 for qualitative results.

2. Ablation Studies

Sensitivity to Number of Training Images We investigate how the number of images in the training set impacts model performance. Figure 2 shows novel views for a varying number of training images. Performance degrades gracefully with a decreasing number of images: While fine detail is reduced significantly, 3D geometry and rigid body motion is preserved.

Sensitivity to Volume Resolution We demonstrate the impact of a smaller volume resolution on model performance. Figure 3 shows novel views for a coarser discretization than the proposed 32 voxels per dimension. While high-frequency detail is degraded, 3D geometry stays consistent.

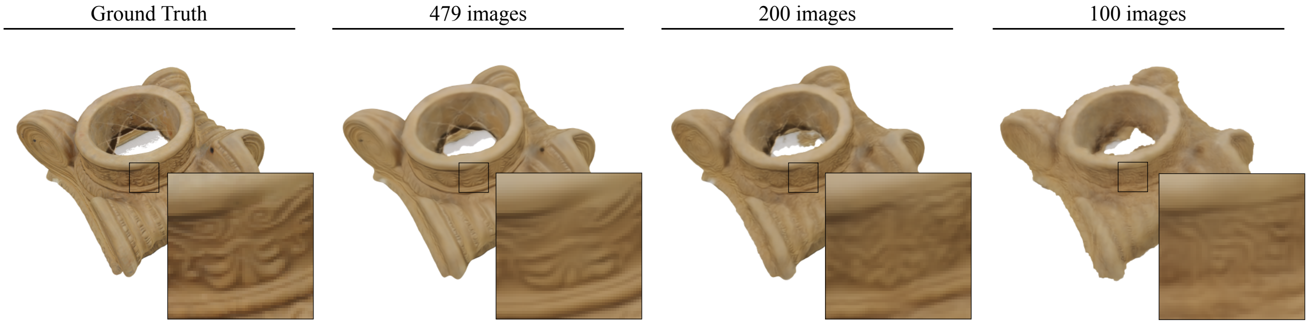

Figure 2: Impact of number of training images on performance. From left to right: ground truth, models trained with 479, 200, and 100 images. While fine detail degrades with a decreasing number of images, the overall geometry stays coherent. Notably, on the pedestal dataset, we still outperform all baselines for 200 images with 30.65dB and with 26.15dB for 100 images, less than a fourth of the data.

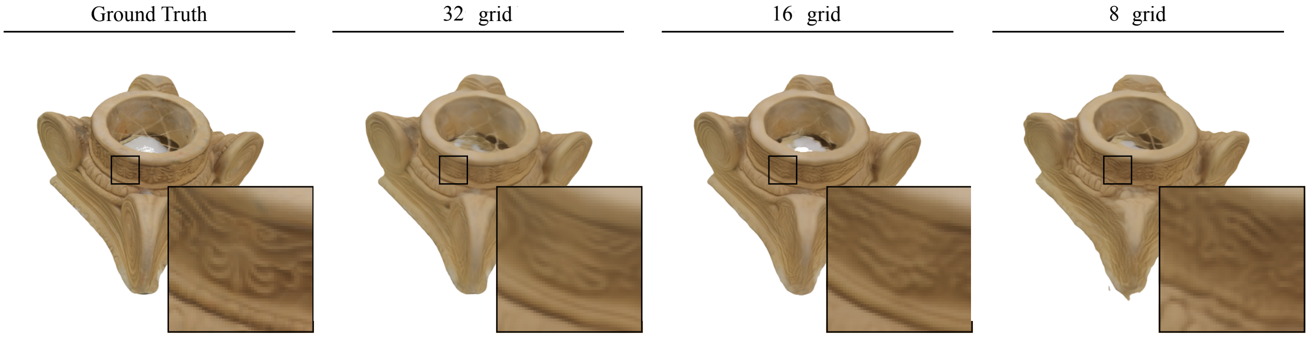

Figure 3: Impact of volume resolution. From left to right: ground truth, models with grid resolution 32, 16 and 8 and PSNRs over the whole test set of 32.35dB, 30.13dB and 25.15dB. Quality deteriorates gracefully, with loss of fine detail but preservation of overall geometry.

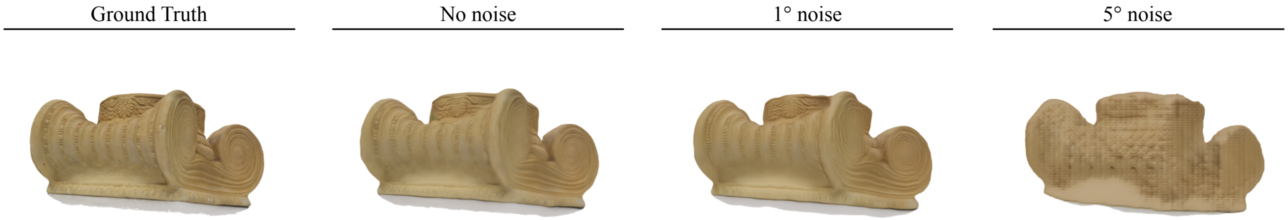

Figure 4: Impact of additive geometric noise in camera poses. From left to right: ground truth, no noise, 1° rotational noise, and 5° rotational noise and PSNRs over the whole test set of 32.35dB, 26.98dB and 23.33dB.

Sensitivity to Additive Rotational Noise We demonstrate the impact of additive uniform random noise added to training poses. We trained the model on the pedestal with 1° and 5° random uniform rotation added to training poses. The model achieves 26.98dB at 1°, still out- performing all baselines, and 23.33dB at 5°, outperforming three baselines. See Fig. 4 for examples. We note that our model has no trouble handling noisy poses obtained via bundle-adjustment.

3. Results on Real-World Captures

Here, we outline additional details on real-world data captured with a digital single-lens reflex camera. For each scene, we captured approximately 450 photographs. We use sparse bundle adjustment to estimate intrinsic and extrinsic camera parameters, as well as a sparse point cloud of keypoints to estimate the scale and center of gravity of the scene. Photographs were subsequently symmetrically center-cropped and downsampled to a resolution of 512 × 512 pixels. Zoom and focus were set at fixed values throughout the capture.

4. Failure Cases

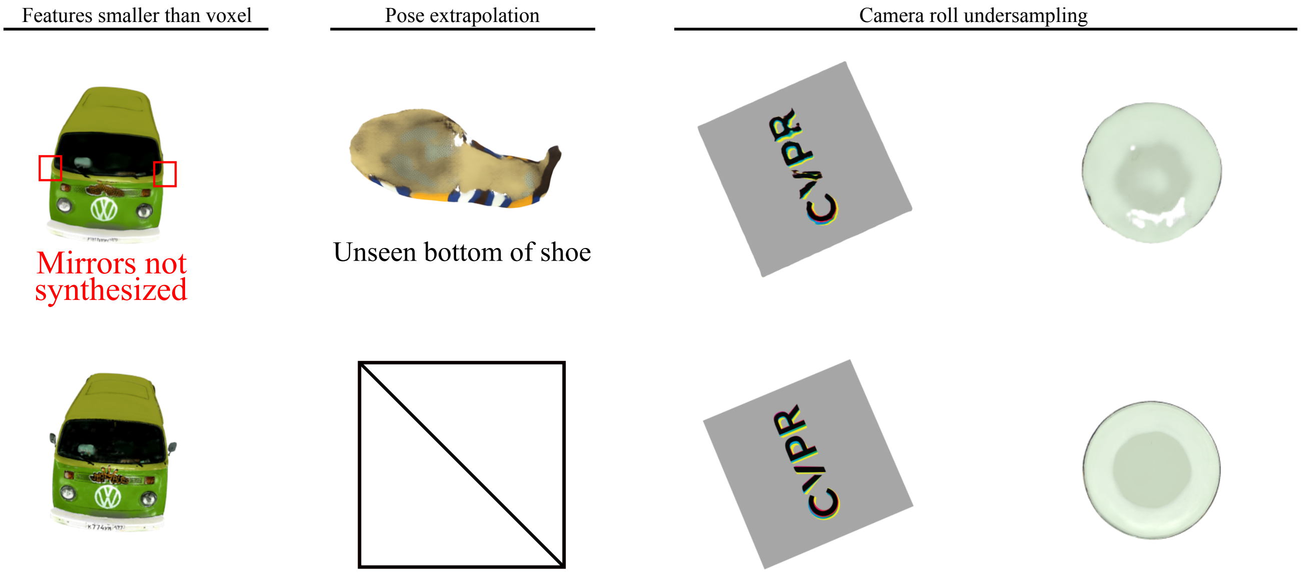

Figure 5: Failure cases of the proposed method. From left to right: If features are significantly smaller than a voxel, our method fails to synthesize them. For strong pose extrapolation, our method may generate imagery with holes or views that are not multi-view consistent. Since our training data does not include variation in camera roll, object views are sampled only sparsely when seen from the top due to the gimbal lock. This may lead to multi-view inconsistencies when objects are seen from the top (see the "V" in the CVPR logo). For the vase, we found this may lead to "holes" in generated images (right) - this may be due to the similarity of the vase color to the background color.

5. DeepVoxels Submodule Architectures

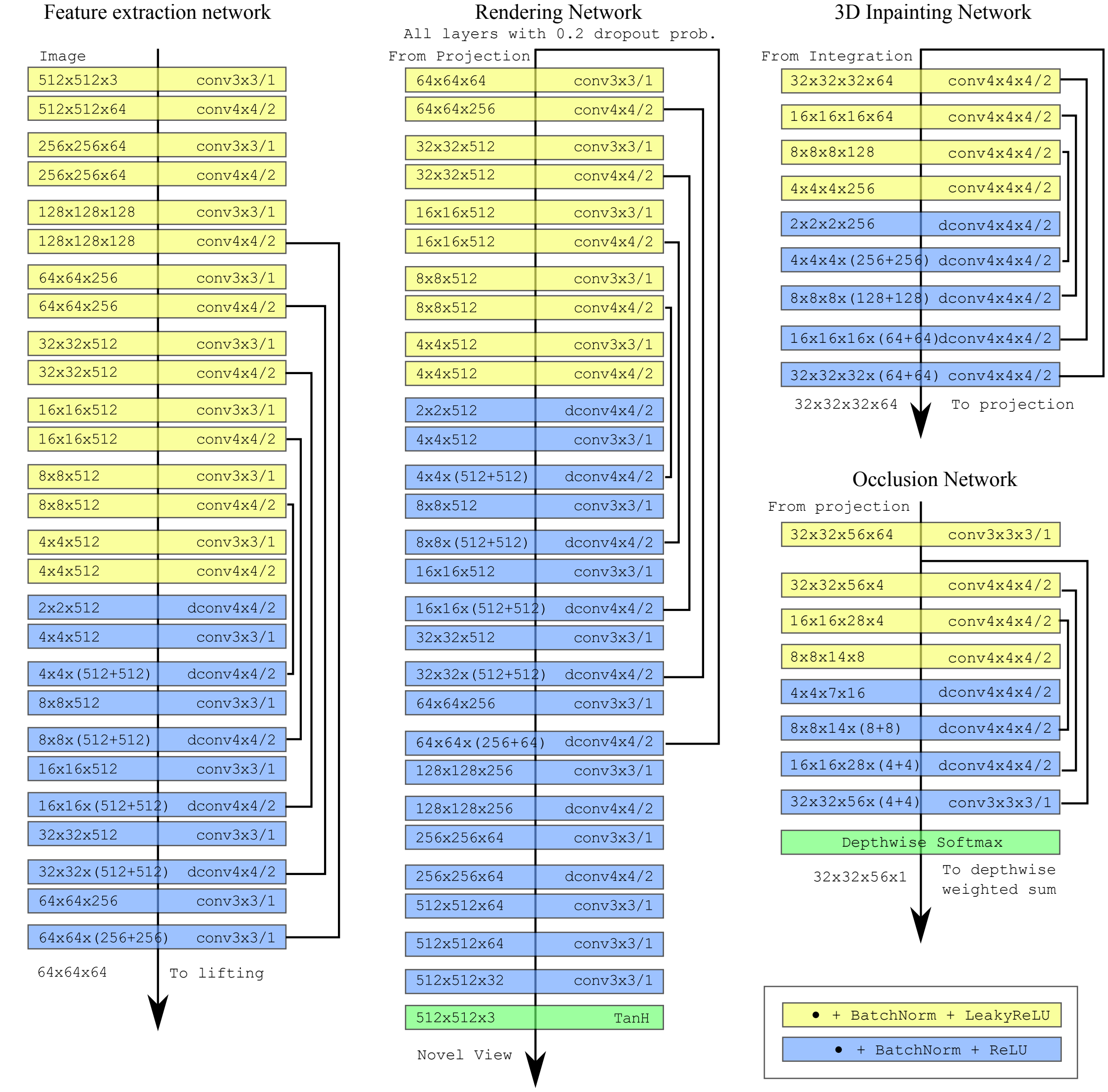

Figure 6: Precise architectures of the feature extraction, rendering, inpainting and occlusion networks. They all follow the basic U-Net structure, while following general best practices in generative network architectures: Reflection padding instead of zero padding, kernel size divisible by stride.

6. Baseline Architecture Tatarchenko et al. [2]

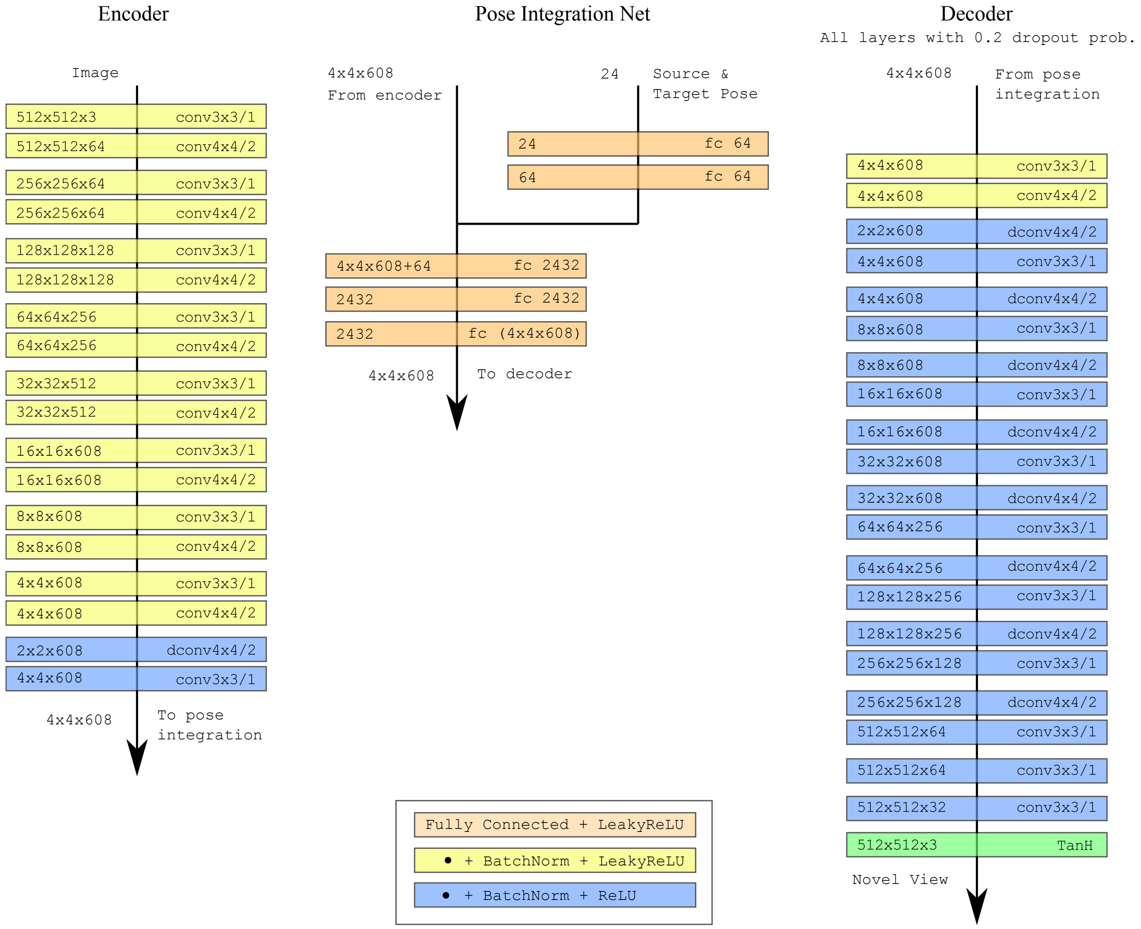

Figure 7: Architectural details of the autoencoder baseline model with latent pose concatenation as proposed by Tatarchenko et al. [2].

7. Baseline Architecture Worrall et al. [3]

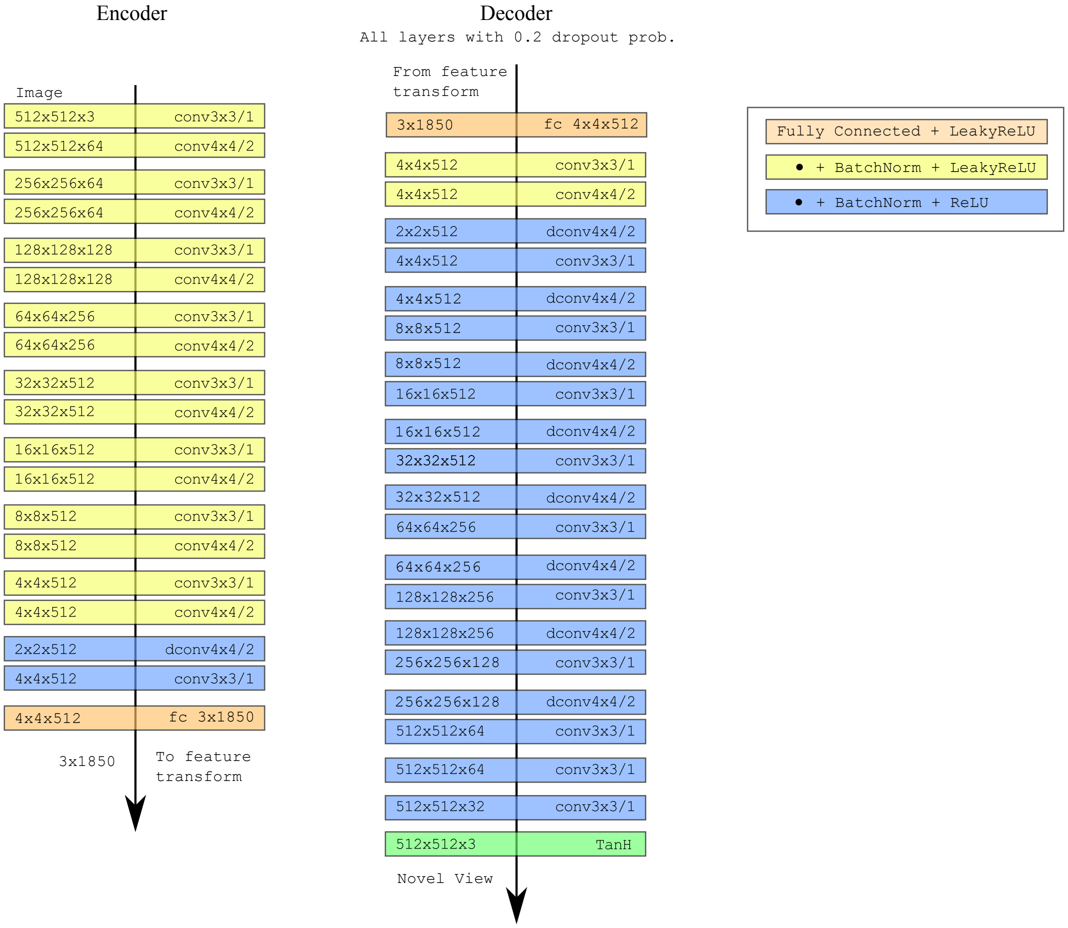

Figure 8: Architectural details of the baseline model based on a rotation-equivariant latent space as proposed by Worrall et al. [3].

8. Baseline Architecture Pix2Pix (Isola et al. [1])

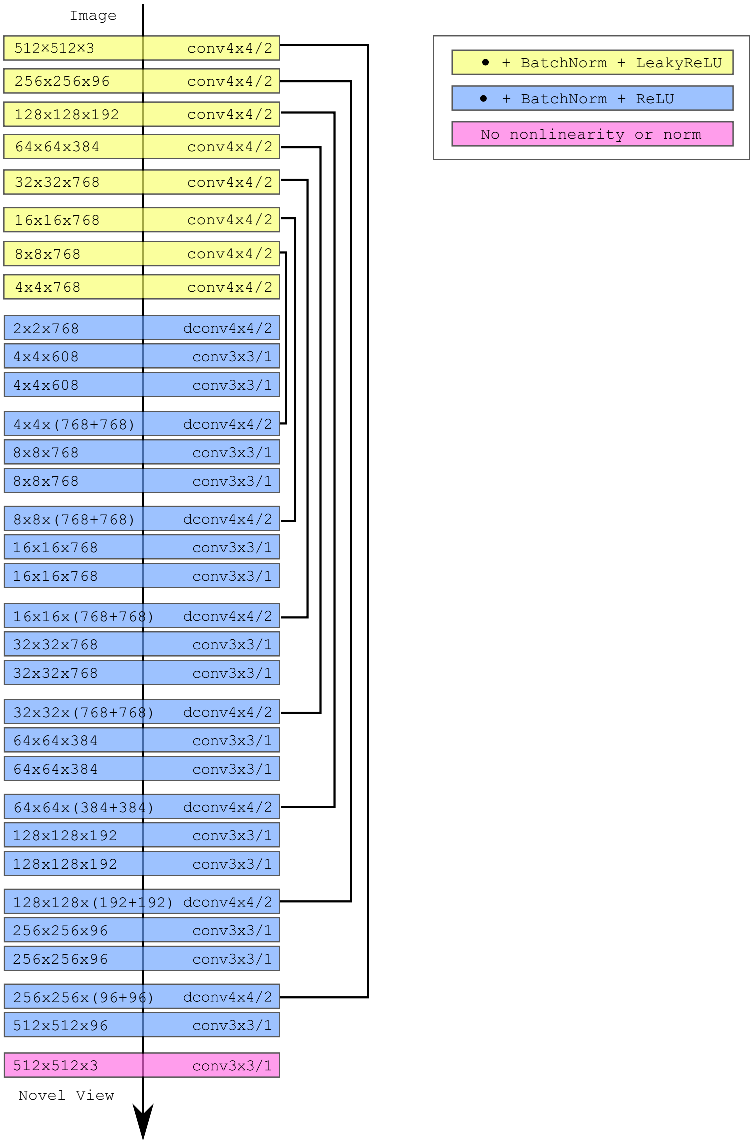

Figure 9: Architectural details of the image-to-image translation baseline model based on Pix2Pix by Isola et al. [1].

9. Comparison of Ground-Truth Depth to Estimated Depth

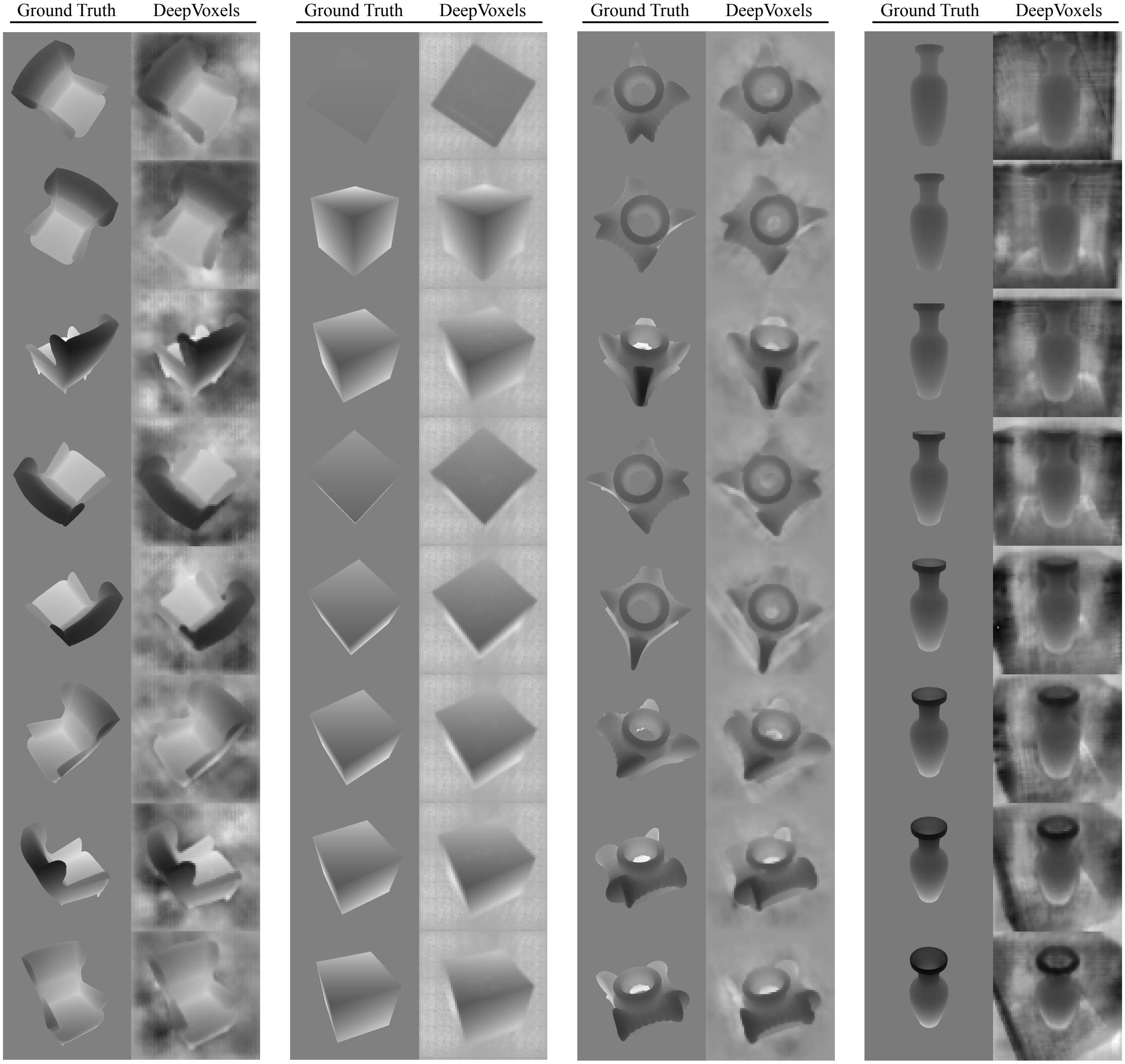

Figure 10: Comparison of ground truth depth maps and the depth maps implicit in the DeepVoxels voxel visibility scores (upsampled from a resolution of 64×64 pixel). We note that these depth maps are learned in a fully unsupervised manner (at no time does our model see a depth map), and only arise out of the necessity to reason about voxel visibility. The background of the depth map is unconstrained in our model, which is why depth values may deviate from ground truth.

10. Pose Extrapolation

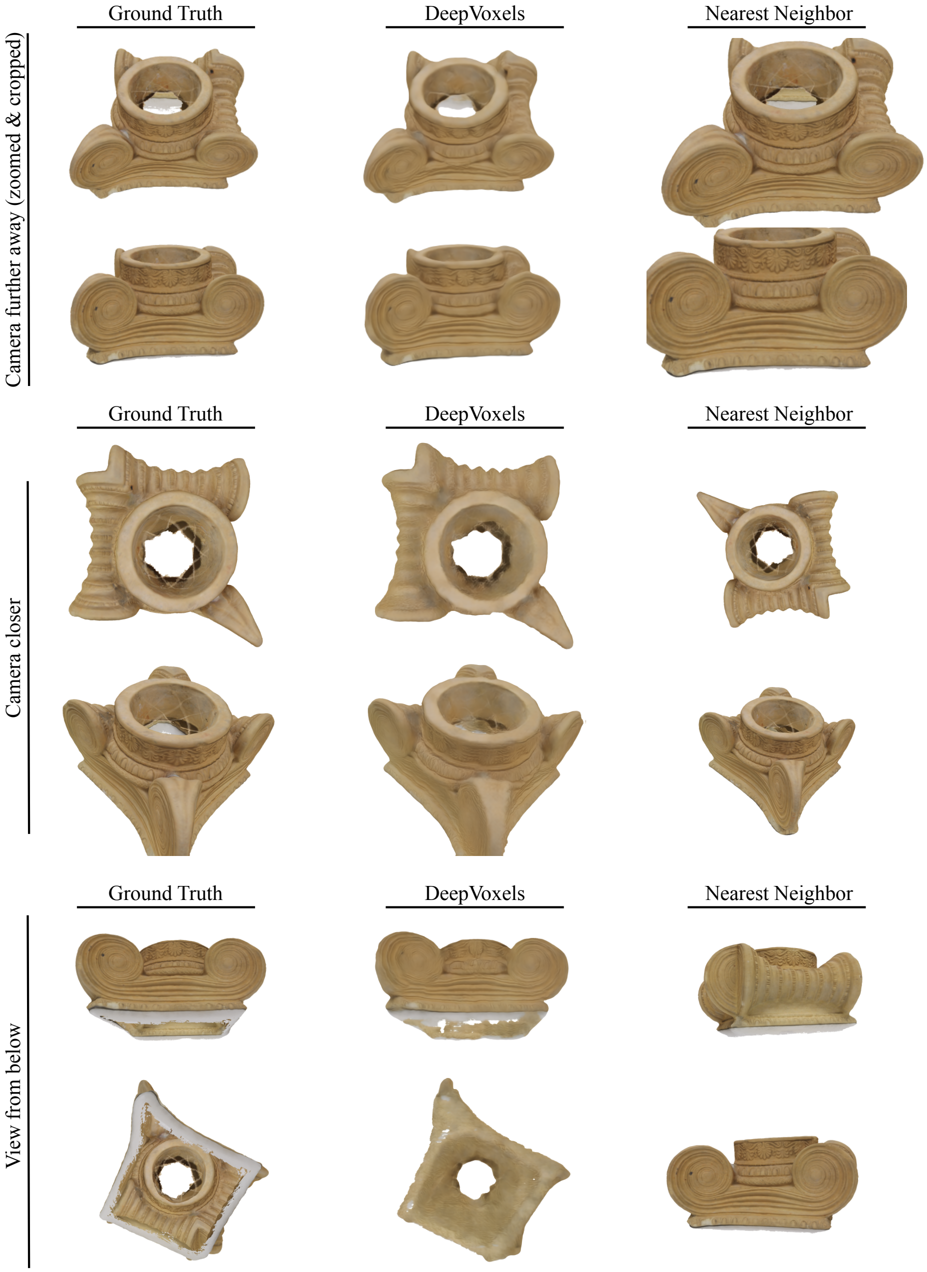

Figure 11: Our training set comprises views sampled at random on the surface of the northern hemisphere. Images in each row are consistently scaled and cropped. We show views that require the model to extrapolate more aggressively - such as increasing the camera distance by a factor of 1.3 (top row), decreasing the camera distance by a factor of 0.75 (middle row) or leaving the northern hemisphere altogether and sampling from the southern hemisphere (bottom row). We show a comparison of ground truth (left column), our model output (center column), and the nearest neighbor in the training set (right column). For the proposed model, detail is lost especially in cases where the model has either never seen these points on the object (bottom row), or where details are seen from closeby for the first time (middle row). Generally, however, the performance degrades gracefully - rigid body motion and general geometry stay consistent, with loss in fine-scale detail and a few failures in occlusion reasoning.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, pages 5967–5976, 2017. 1, 7

[2] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Single-view to multi-view: Reconstructing unseen views with a convolutional network. *CoRR abs/1511.06702*, 1(2):2, 2015. 1, 5

[3] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Interpretable transformations with encoder-decoder networks. In *Proc. ICCV*, volume 4, 2017. 1, 6